

Challenges in Kurdish Text Processing

Kyumars Sheykh Esmaili

Nanyang Technological University
N4-B2a-02, 639798 Singapore
kyumarss@ntu.edu.sg

Abstract. Despite having a large number of speakers, the Kurdish language is among the less-resourced languages. In this work we highlight the challenges and problems in providing the required tools and techniques for processing texts written in Kurdish. From a high-level perspective, the main challenges are: the inherent diversity of the language, standardization and segmentation issues, and the lack of language resources.

1 Introduction

Kurdish language belongs to the Indo-Iranian family of Indo-European languages. Its closest better-known relative is Persian. Kurdish is spoken in Kurdistan, a large geographical area spanning the intersections of Iran, Iraq, Turkey, and Syria. It is one of the two official languages in Iraq and has a regional status in Iran.

Despite having 20 to 30 millions of speakers¹, Kurdish is among the less-resourced languages and there are very few tailor-made tools available for processing texts written in this language. Similarly, it has not seen much attention from the IR and NLP research communities and the existing literature can be summarized as a few attempts in building corpus and lexicon for Kurdish [2,8]. In order to provide the basic tools and techniques for processing Kurdish, we have recently launched a project at University of Kurdistan (UoK²). This paper gives an overview of the main challenges that we need to address throughout this project.

Before proceeding to enumerate the challenges, we would first like to highlight a few things about the scope and methodology of the current paper. Firstly, in this work we only consider the two largest and closely-related branches of Kurdish –namely Kurmanji (or Northern Kurdish) and Sorani (or Central Kurdish)– and exclude the other smaller and distant dialects. Secondly, in the interest of space, we give greater importance to the issues that are specific to Kurdish and refer to the related papers for in-depth discussion of issues that are shared between Kurdish and other languages (in particular, Persian, Arabic, and Urdu). Finally, we restrict our discussion to the bag-of-words (i.e., IR) model and do not address the structural (i.e., NLP) aspects.

¹ Numbers vary, depending on the source.

² Project’s website: <http://eng.uok.ac.ir/esmaili/research/klpp/en/main.htm>.

2 Challenges

We have categorized the main challenges into five groups. While the first two groups are concerned with the diversity aspect of the Kurdish language, the third and fourth highlight the processing difficulties and the last one examines the depth of resource-scarcity for Kurdish.

2.1 Dialect Diversity

The first and foremost challenge in processing Kurdish texts is its dialect diversity. In this paper we focus on Kurmanji and Sorani which are the two most important Kurdish dialects in terms of number of speakers and degree of standardization [3]. Together, they account for more than 75% of native Kurdish speakers [8].

The features distinguishing these two dialects are mainly morphological (the phonological differences are explained in the next section). The important morphological differences are [5,3]:

- Kurmanji is more conservative in retaining both gender (feminine:male) and case opposition (absolute:oblique) for nouns and pronouns. Sorani has largely abandoned this system and uses the pronominal suffixes to take over the functions of the cases³.
- in the past-tense transitive verbs, Kurmanji has the full ergative alignment but Sorani, having lost the oblique pronouns, resorts to pronominal enclitics.
- in Sorani, passive and causative can be created exclusively via verb morphology, in Kurmanji they can also be formed with the verbs **hatin** (to come) and **dan** (to give) respectively.
- the definite suffix **-eke** appears only in Sorani

2.2 Script Diversity

Due to geopolitical reasons, each of the two aforementioned dialects has been using its own writing system. In fact, Kurdish is considered a *bi-standard* language [2], with Kurmanji written in Latin-based letters and Sorani written in Arabic-based letters. Both of these systems are almost totally phonetic [2]. As noted before, Sorani and Kurmanji are not morphologically identical and since these systems reflect the phonology of their corresponding dialects, there is no bijective mapping between them. In Figure 1 we have included three tables to demonstrate the non-trivial mappings between these two writing systems. It should be noted that the table in c) contains approximate equivalences. These mappings are in line with the list proposed in [4].

2.3 Normalization

The Unicode assignments of the Arabic-based Kurdish alphabet has two potential sources of ambiguity which should be dealt with carefully:

³ Although there is evidence of gender distinctions weakening in some varieties of Kurmanji [3].

Unicode Value	Latin based Char.	Arabic based Char.	Unicode Value
0048	H	ه	06BE
0068	h		
0049	I	ـ	-
0069	i		
0055	U	و	0648
0075	u	و	0648
0057	W	و	
0077	w	و	06CC
0059	Y	ی	
0079	y	ی	

Unicode Value	Arabic based Char.	Latin based Char.	Unicode Value
0648	و	U	0055
		u	0075
		W	0057
		w	0077
06BE	ه	H	0048
		h	0068
		E	0045
		e	00EA
06CC	ی	İ	00CE
		ı	00EE
		Y	0059
		y	0079

Unicode Value	Arabic based Char.	Latin based Char.	Latin based Char. (approx.)	Unicode Value
0626	ئ	-	-	-
062D	ژ	-	H	0048
			h	0068
0695	ڕ	-	R+R	0052
0639	ع	-	r+r	0072
			E	0045
			e	00EA
063A	غ	-	X	0058
			x	0078
06B5	ل	-	L+L	004C
			l+l	006C

a) From Latin-based to Arabic-based **b)** From Arabic-based to Latin-based **c)** From Arabic-based to Latin-based (approx.)

Fig. 1. Non-Trivial Mappings between Arabic-based and Latin-based Kurdish Alphabets.

- for some letters such as **ye** and **ka** there are more than one Unicode [7]. During the normalization phase, the occurrences of these multi-code letters should be unified.
- as in Urdu, the isolated and final forms of the Arabic letter **ha** constitute one letter (pronounced e), whereas the initial and medial forms of the same Arabic letter constitute another letter (pronounced h), for which a different Unicode encoding is available [8,2]. In many electronic texts, these letters are written using only the **ha**, differentiated by using the *zero-width non-joiner* character that prevents a character from being joined to its follower. This distinction must be taken into account in the normalization phase.

2.4 Segmentation and Tokenization

Segmentation refers to the process of recognizing boundaries of text constituents, including sentences, phrases and words. Compared to Persian and Arabic, this process is relatively easier in Kurdish, mainly because short vowels are explicitly represented in the Kurdish writing systems.

In fact, as discussed in [1], the absence of short vowels contributes most significantly to ambiguity in Arabic language, causing difficulty in homograph resolution, word sense disambiguation, part-of-speech detection. In Persian, its negative consequence is more visible in detecting the *Izafe* constructs [7]⁴.

Despite incorporating short vowels, the Arabic-based Kurdish alphabet still suffers from two problems which are inherited from the Arabic writing system:

- Arabic alphabet does not have capitalization and therefore it is more difficult to recognize sentence boundaries as well as recognizing Named Entities.
- Space is not a deterministic delimiter and boundary sign [7]. It may appear within a word or between words, or may be absent between some sequential words. There are some proposals on how to tackle this issue in Persian [6] and Urdu [9].

⁴ *Izafe* is an unstressed vocal **-e** or **-i** added between prepositions, nouns and adjectives in a phrase. It approximately corresponds to the English preposition *of*. This construct is frequently used in both Persian and Kurdish languages.

2.5 Lack of Language Resources

Kurdish is a resource-scarce language for which the only linguistic resource available on the Web is raw text [8].

More concretely, in spite the few attempts in building corpus [2] and lexicon [8], Kurdish still does not have any large-scale and reliable general/domain-specific corpus. Furthermore, no test collection –which is essential in evaluation of Information Retrieval systems– or stemming algorithm has been developed for Kurdish so far.

Lastly, although Kurdish is well served with dictionaries [3], it still lacks a WordNet-like semantic lexicon.

3 Conclusions

Kurdish text processing poses a range of challenges. The most important one is the dialect/script diversity which has resulted in a *bi-standard* situation. As the examples in [2] show, the “same” word, when going from Kurmanji to Sorani, may at the same time go through several levels of change: writing systems, phonology, morphology, and sometimes semantics. This clearly shows that the mapping between these two dialects are more than *transliteration*, though less complicated than *translation*. Any text processing system designed for Kurdish language should develop and exploit a mapping between these two standards.

On the technical side, providing the required processing techniques –through leveraging the existing techniques or designing new ones if needed– offers many avenues for future work. However, as a critical prerequisite to most of these tasks, a core set of language resources must be available first. At UoK, we have taken the first step and are currently working on building a large standard test collection for Kurdish language.

References

1. A. Farghaly and K. F. Shaalan. Arabic Natural Language Processing: Challenges and Solutions. *ACM Trans. on Asian Lang. Info. Processing*, 8(4), 2009.
2. Gérard Gautier. Building a Kurdish Language Corpus: An Overview of the Technical Problems. In *Proceedings of ICEMCO*, 1998.
3. Geoffrey Haig and Yaron Matras. Kurdish Linguistics: A Brief Overview. *Sprachtypologie und Universalienforschung / Language Typology and Universals*, 55(1), 2002.
4. Amir Hassanpour. *Nationalism and Language in Kurdistan, 1918-1985*. Mellen Research University Press, 1992.
5. David N. MacKenzie. *Kurdish Dialect Studies*. Oxford University Press, 1961.
6. M. Shamsfard, H.Sadat Jafari, and M. Ilbeygi. STeP-1: A Set of Fundamental Tools for Persian Text Processing. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, 2010.
7. Mehrnoush Shamsfard. Challenges and Open Problems in Persian Text Processing. In *Proceedings of the 5th Language and Technology Conference (LTC)*, 2011.

8. G  r  ldine Walther and Beno  t Sagot. Developing a Large-scale Lexicon for a Less-Resourced Language. In *SaLTMiL's Workshop on Less-resourced Languages (LREC)*, 2010.
9. U. I. Bajwa Z. Rehman, W. Anwar. Challenges in Urdu Text Tokenization and Sentence Boundary Disambiguation. In *Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP)*, 2011.